Towards robust cross-linguistic comparisons of phonological networks

Towards robust cross-linguistic comparisons of phonological networks

Philippa Shoemark
School of Informatics
University of Edinburgh
p.j.shoemark@sms.ed.ac.uk

Sharon Goldwater
School of Informatics
University of Edinburgh
sgwater@inf.ed.ac.uk

James Kirby
Linguistics and English Language
University of Edinburgh
j.kirby@ed.ac.uk

Rik Sarkar
School of Informatics
University of Edinburgh
rsarkar@inf.ed.ac.uk

Abstract

Recent work has proposed using network science to analyse the structure of the mental lexicon by viewing words as nodes in a *phonological network*, with edges connecting words that differ by a single phoneme. Comparing the structure of phonological networks across different languages could provide insights into linguistic typology and the cognitive pressures that shape language acquisition, evolution, and processing. However, previous studies have not considered how statistics gathered from these networks are affected by factors such as lexicon size and the distribution of word lengths. We show that these factors can substantially affect the statistics of a phonological network and propose a new method for making more robust comparisons. We then analyse eight languages, finding many commonalities but also some qualitative differences in their lexicon structure.

1 Introduction

Studies suggest that the ease with which a word is recognised or produced is affected by the word's phonological similarity to other words in the mental lexicon (often operationalised as *neighbourhood density*, i.e., the number of words that differ from a target word by just a single phoneme) (Luce and Pisoni, 1998; Harley and Bown, 1998; Vitevitch, 2002; Ziegler et al., 2003). Yet the nature of these effects is not always consistent between languages. For example, in English, the neighbourhood density of a word was found to correlate positively with reaction times in a picture naming task (Vitevitch, 2002), and negatively with the speed and accuracy of participants' responses in a lexical decision task (Luce and Pisoni, 1998). However, in Spanish the opposite pattern was found: words with higher neighbourhood densities were produced less quickly in a picture naming task (Vitevitch and Stamer, 2006), and recognised more quickly and accurately in an auditory lexical decision task (Vitevitch and Rodríguez, 2005).

A possible explanation for such cross-linguistic variation is that the different effects of neighbourhood density (a *local* measure of lexicon structure) might result from differences in the *global* lexicon structure: for example, if one language exhibits much greater similarity between words overall, this could affect how language processing mechanisms develop, leading to qualitatively different behaviour.

One way to analyse the global phonological structure of the mental lexicon is by representing it as a *phonological network* (Vitevitch, 2008): a graph in which nodes correspond to the wordforms in a lexicon, and edges link nodes which are phonologically similar according to some metric (typically, words which differ by exactly one phoneme, i.e. they have a Levenshtein distance of one). The structure of the network can then be analysed quantitatively using measures from network science. While neighbourhood density (i.e., a node's degree) is one local measure of connectivity, other measures can better capture the global structure of the network. By comparing these measures across languages, we might find explanations for the behavioural differences mentioned above.

As well as providing insight into cross-linguistic variations in language processing, cross-linguistic comparisons of phonological network structure could also uncover universal properties of language

or typological generalisations. Indeed, Arbesman et al. (2010) argued based on an analysis of phonological networks from five languages that these networks share several structural properties that distinguish them from other naturally occurring networks, suggesting some important underlying organisation. Though specific hypotheses were not presented in this work, one of the authors suggested in an earlier analysis of the English lexicon that such properties might arise due to particular learning processes (Vitevitch, 2008).

However, work by Gruenenfelder and Pisoni (2009) found that several of the structural properties discussed above were also found in a random pseudolexicon with the same word lengths as a real English lexicon, but with the phonemes in each word chosen at random. Thus, they argued that these structural properties are simply a by-product of the way phonological networks are defined (by connecting similar strings) and should not be taken as evidence of particular growth processes (or presumably, any other cognitive pressures Arbesman et al. might later have had in mind).

These studies highlight some important methodological issues that need to be resolved if we hope to use network analysis as a tool for cross-linguistic studies. In order to make meaningful comparisons between different languages' networks, we need to determine what constitutes a large or small difference in phonological network statistics by comparing *all* languages to appropriate random baselines. In addition, there are two other factors that have not been explicitly considered in previous studies of phonological networks. First, we don't know how the *size* of a phonological network (number of nodes) affects its statistics. The lexicons in Arbesman et al.'s study ranged from 2500 words (Hawaiian) to 122,000 words (Spanish), yet if the size of the Spanish lexicon had also been 2500 words, it might have yielded quite different statistics. Second, there is a lack of consensus about whether phonological networks should be constructed from *lemmas* or from *all wordforms* including morphological variants, and in some cases, datasets may only be available for one or the other option. Therefore, we need to understand how including or excluding inflectional variants affects the measured statistics of phonological networks.

In this paper, we investigate the questions above, synthesizing arguments from the literature with our own analyses to propose a new method for comparing phonological networks across languages. Using this method, we compare network statistics across eight different languages to appropriate baselines at a range of lexicon sizes. We show that Gruenenfelder and Pisoni's (2009) findings for English extend to the other seven languages we consider, supporting their argument that the small-world properties of phonological networks should not be used as evidence of particular cognitive/growth processes. We also find that network statistics vary with lexicon size within each language, but not always in the same way. These differences provide a first step in investigating the relationship between cross-linguistic variation in language processing and global lexicon structure.

2 Background

The use of network science to study the phonological structure of the lexicon was first proposed by Vitevitch (2008). He and later authors converged on using several standard measures from network science, which we will also employ. These are:

Degree assortativity coefficient In some networks, nodes tend to connect to other nodes that have similar *degrees* (numbers of neighbours) to their own. The extent to which a network exhibits this property can be quantified using the *degree assortativity coefficient*, which is defined as the Pearson correlation coefficient r of the degrees of the nodes at either end of each edge. So, r lies between -1 (the higher a node's degree is, the lower the degrees of its neighbours) and 1 (nodes connect only to other nodes of the same degree), with $r = 0$ if there is no correlation between the degrees of neighbouring nodes.

Networks with positive degree assortativity are relatively robust. Empirical studies show that in such networks many nodes can be removed without substantially reducing their connectivity (Newman, 2003).

Fraction of nodes in the giant component Complex networks often have many distinct connected components. Often, a single *giant component* contains a much larger fraction of the nodes than any other component, and this fraction helps characterise the global connectivity of the network.

Average shortest path length (ASPL) The *shortest path length* between two nodes v and w, which we denote $d(v, w)$, is the minimum number of edges that must be traversed to get from node

v to node w. The ASPL is then the mean of the shortest path lengths between all pairs of nodes, and is given by the equation

$$\text{ASPL} = \sum_{v,w \in V} \frac{d(v,w)}{|V|(|V|-1)},$$

where V denotes the set of all nodes in the network. Since paths do not exist between mutually disconnected components, there are different ways to compute ASPL for graphs with disconnected components; all values reported in this paper compute the average across all pairs of nodes in the giant component only.

Average clustering coefficient A node's *clustering coefficient* measures the 'cliquishness' of its neighbourhood, and is defined as the number of edges that exist between its neighbours divided by the number of possible edges between them:

$$C(v) = \frac{2|\{\, e_{u,w} \in E : e_{v,u} \in E, e_{v,w} \in E\}|}{k(v)(k(v)-1)},$$

where E denotes the set of all edges in the network, $e_{x,y}$ denotes an edge between nodes x and y, and $k(x)$ denotes the degree of node x. The clustering coefficient is undefined for nodes with $k < 2$, since the denominator reduces to zero for such nodes. We report the mean clustering coefficient over all nodes in the giant component; nodes with fewer than two neighbours are assigned a coefficient of zero.[1]

A word's clustering coefficient has been found to predict behavioural measures in both lexical access (Chan and Vitevitch, 2009) and adult and child word learning (Goldstein and Vitevitch, 2014; Carlson et al., 2014).

Small-world property Small world networks (Watts and Strogatz, 1998) are characterized by short ASPL relative to their size and high average clustering coefficients relative to what one would expect from an equivalent *Erdős-Rényi graph*—one with the same number of nodes and edges as the real graph, but where edges are placed randomly between pairs of nodes. A distinctive property of these networks is their easy searchability: it is usually possible to find short paths between nodes in a decentralized fashion using only small quantities of information per node when the network admits embedding in a suitable space (Kleinberg, 2000; Sarkar et al., 2013). It has been suggested that easy searchability could be relevant for spreading-activation models of lexical processing (Chan and Vitevitch, 2009) and in lexical acquisition (Carlson et al., 2011).

Using the measures above, Vitevitch (2008) analysed a lexicon of English, and Arbesman et al. (2010) extended the analysis to five lexicons representing languages from different language families. They found several characteristics common to these networks. All five lexicons were found to exhibit the small-world property, having similar ASPLs to those expected in comparable Erdős-Rényi graphs, but average clustering coefficients that were several orders of magnitude larger. The phonological networks were also marked by high degree assortativity, with coefficients ranging from 0.56 to 0.76, in contrast to typical values of 0.1 to 0.3 for social networks, and -0.1 to -0.2 for biological and technical networks. The giant components in the phonological networks all contained less than 70% of nodes (in three cases, less than 40%), whereas the giant components of social, biological, and technical networks typically contain 80-90% of nodes. Arbesman et al. suggested that "together, these observed characteristics hint at some deeper organization within language" (2010: 683).

Nevertheless, Arbesman et al. also found some quantitative variation in the phonological network statistics across languages—for example, the Mandarin network had an ASPL almost twice that of the Hawaiian network, a clustering coefficient twice that of the Spanish network, and the fraction of nodes in its giant component was almost twice that of the English network. However, we don't know if these differences are meaningful, since the expected variability of these statistics in phonological networks has not been established. In addition, since the lexicon sizes varied widely across languages, the differences in network statistics may have been due to this variation rather than to more interesting differences between the languages.

Gruenenfelder and Pisoni (2009) started to address these issues by considering a random baseline network for English. They constructed a pseudolexicon by randomly generating phoneme sequences with the same lengths as the words in an English lexicon[2], and found that the phonological network

[1] Some researchers instead define the coefficient for such nodes to be one, whilst others exclude such nodes from the average (Schank and Wagner, 2004).

[2] Both the English lexicon and the pseudolexicon were limited to words of only 2 to 5 phonemes in length.

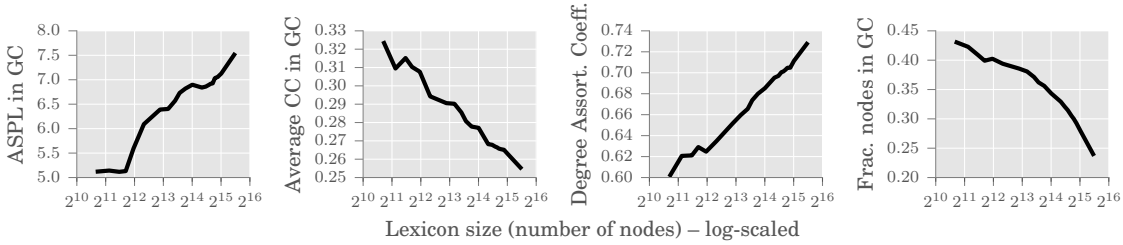

Figure 1: Phonological network statistics of English lemmas, as a function of lexicon size.

of this random pseudolexicon also exhibited the small-world property and a high degree assortativity coefficient. They concluded that these characteristics are likely to occur in any network whose nodes represent random sequences of parts (i.e., phonemes), and whose edges are determined by the overlap of these parts. Thus they argued that high assortative mixing by degree and small-world characteristics should not be taken to indicate a deeper organisational principle in the lexicon.

So, although Gruenenfelder and Pisoni (2009) have analysed some properties of a randomly generated lexicon with the same size, word-lengths, and phoneme inventory as a particular English lexicon, it remains to be seen whether these properties are characteristic of randomly generated lexicons in general, or to what extent they vary across different lexicon sizes, word-length distributions, and phoneme inventory sizes. In the following sections we show that all of these factors affect the statistics of both random and real lexicons; we then propose a more robust method for making cross-linguistic comparisons of phonological networks. While our method is still not enough to draw strong quantitative conclusions in all cases, we are able to shed further light on a number of the claims and questions raised above, and we also discover some cross-linguistic differences in lexicon structure that warrant further investigation.

3 Effect of lexicon size

We begin by asking whether the size of a phonological network may affect its statistics. For this analysis, we use only a single language (English).

3.1 Method

We start with the 44,841 English lemmas in the CELEX database (Baayen et al., 1995), which includes both word frequencies and phonemic transcriptions. We derive from this original lexicon a series of sublexicons of decreasing sizes, by progressively filtering out batches of words with the lowest frequencies. Thresholding a lexicon by frequency simulates drawing a lexicon from a smaller corpus or dictionary, since the more frequently a word is used, the more likely it is to appear in even a small corpus or dictionary. For each (sub)lexicon, we associate each distinct phonological form with a unique node and place edges between pairs of nodes that differ by one phoneme (insertion, deletion, or substitution). To construct the networks and compute their statistics we use the NetworkX Python package (Hagberg et al., 2008).

3.2 Results and discussion

Figure 1 shows the values of four network statistics as a function of lexicon size. All the values fall within the range found across languages by Arbesman et al. (2010). However, all four statistics do vary with lexicon size, suggesting that comparisons between networks should only be made between lexicons of similar size.

One way to attempt such quantitative comparisons across languages could be to subsample from each lexicon in order to obtain lexicons of the same size. However, we don't know if the slopes of these plots will be the same across languages. Consider the hypothetical plot in Figure 2:

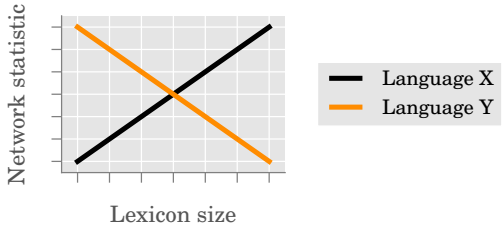

Figure 2: Hypothetical scenario where the value of a phonological network statistic is positively correlated with lexicon size in Language X, but negatively correlated in Language Y.

In this case, controlling for lexicon size is not enough, since if we choose a small lexicon size then Language X will have a smaller statistic than Language Y, whereas if we choose a large size then

the opposite holds. This observation motivates us to compare statistics across a *range* of sizes rather than using point estimates as in previous work.

Indeed, an established technique for comparing the properties of real networks against random baselines is to plot the value of a network statistic as a function of network size, and compare the slope obtained for random networks against the trend observed in real networks (Albert and Barabási, 2002). However, care must be taken in choosing appropriate random baselines for phonological networks, due to the issues described next.

4 Effects of word-length distribution and phoneme inventory size

Recently, Stella and Brede (2015) pointed out that due to the way in which nodes and edges are typically defined in phonological networks, the statistics of such networks are highly sensitive to the distribution of word lengths in the lexicon, and to a lesser extent, the size of the phoneme inventory. Stella and Brede considered the set of all possible 'words' (i.e. possible sequences of phonemes) that could be formed using a given phoneme inventory, and noted that the number of possible n-phoneme words scales exponentially with n, while the number of possible neighbours of an n-phoneme word scales linearly with n. Hence, if we randomly sample a pair of words from the set of all possible words, then the shorter their lengths are, the more likely it is that the sampled words will be neighbours. Thus, lexicons with a higher proportion of short words will tend to be more densely connected, regardless of any other phonological properties. Also, since the number of possible n-phoneme words scales faster with the size of the phoneme inventory than does the number of possible neighbours of an n-phoneme word, we expect the size of the phoneme inventory to affect the connectivity of a phonological network, albeit by a smaller factor than the distribution of word lengths.

Unlike lexicon size, the word-length distribution and phoneme inventory size are inherent properties of a language, so these confounds make it difficult to directly compare network statistics across languages, even after controlling for lexicon size. In making cross-linguistic comparisions, we would like to be able to identify differences between languages beyond the fact that their lexicons have different word length distributions.

Therefore, rather than directly comparing the statistics of real lexicons across languages, we propose to generate separate pseudolexicons for each language that match the word-length distribution and phoneme inventory size of that language. We can then examine the differences between these pseudolexicons and the real lexicons over a range of lexicon sizes, and compare these differences across languages. Using this method we can better evaluate some of the claims made by previous authors and reveal some previously undetected variation in network structure across languages.

5 Cross-linguistic comparison

5.1 Data and method

We analyse phonological networks from eight different languages: English, Dutch, German, French, Spanish, Portuguese, Polish, and Basque. Where possible, we have obtained for each language a lexicon consisting only of lemmas, and another with separate entries for phonemically distinct inflectional variants.[3] Each lexical entry consists of a phonemically transcribed word and a corpus-derived estimate of its frequency. The sources and sizes of the lexicons are listed in Table 1. From each of these original lexicons, we derive a series of sublexicons of decreasing sizes, by progressively filtering out batches of low-frequency words.

For each real lexicon and derived sublexicon, we generate 20 random pseudolexicons with the same size, phoneme inventory size, and word-length distribution[4]. For each lexicon size in each language, we compute the mean and standard deviation of each statistic across the 20 pseudolexicons, as well as the statistics for the comparable real (sub)lexicon.

5.2 Results and discussion

We first consider how the average word length varies across our sample lexicons. Figure 3 shows that average word lengths vary with lexicon size (tending to increase as more infrequent words are included in the lexicon), as well as across languages (average word lengths in English and French are substantially shorter than in Spanish).

[3] We were unable to obtain phonemic transcriptions for Portuguese inflected wordforms, or reliable frequencies for Spanish lemmas.

[4] Specifically, we replicate Gruenenfelder and Pisoni's procedure for generating their 'Word Length Only' lexicon, except that we match the entire word-length distribution, not just the number of two-, three-, four-, and five-segment words.

Language	Lexicon Type	Size	Source of pronunciations	Source of frequencies
English	Lemmas	44,841	CELEX (Baayen et al., 1995)	CELEX
	All wordforms	87,263	CELEX	CELEX
Dutch	Lemmas	117,048	CELEX	CELEX
	All wordforms	300,090	CELEX	CELEX
German	Lemmas	50,481	CELEX	CELEX
	All wordforms	353,679	CELEX	CELEX
French	Lemmas	43,361	Lexique (New et al., 2001)	Lexique
	All wordforms	71,334	Lexique	Lexique
Portuguese	Lemmas	18,656	Porlex (Gomes and Castro, 2003)	CORLEX (Bacelar do Nascimento, 2003)
Spanish	All wordforms	42,461	CALLHOME (Garrett et al., 1996)	CALLHOME
Polish	Lemmas	6024	GlobalPhone (Schultz, 2002)	SUBTLEX-PL (Mandera et al., 2014)
	All wordforms	25,623	GlobalPhone	SUBTLEX-PL
Basque	Lemmas	9102	E-hitz (Perea et al., 2006)	E-hitz
	All wordforms	99,491	E-hitz	E-hitz

Table 1: Sources and sizes of lexicons. Sizes refer to the number of distinct *phonological* forms: sets of words which have distinct spellings and/or senses but the same phonemic transcription are conflated into a single phonological wordform.

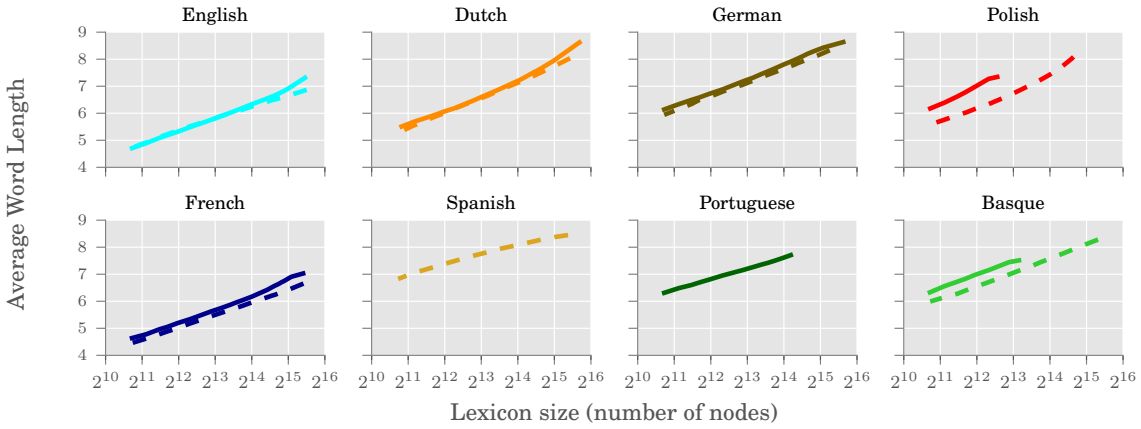

Figure 3: Average word length as a function of lexicon size. Solid lines are for lemmas; dashed lines are for all wordforms.

Results for the four measures of network structure and connectivity defined above are presented in Figure 4. We first discuss the similarities between different languages, focusing on the claims made by previous researchers; we then discuss some cross-linguistic differences.

5.2.1 Cross-linguistic similarities

As noted by Arbesman et al. (2010), there are some striking similarities across the languages, especially relative to other types of networks. To test for the small-word behaviour that Arbesman et al. found in their networks, we computed estimates of the ASPLs and clustering coefficients of Erdős-Rényi graphs matched to the giant components of each of our lexicons.[5] The ASPLs all ranged from 4 to 6, which is somewhat smaller than in our real lexicons, but considered similar according to the conventions used to test for the small-world property (Watts and Strogatz, 1998). The clustering coefficients of the Erdős-Rényi graphs ranged between 0.0003 and 0.03, orders of magnitude smaller than the values of 0.17 to 0.37 for our real lexicons; again, according to the usual conventions (Watts and Strogatz, 1998), these results indicate that all of our real lexicons exhibit the small-world property.

However, all of our pseudolexicons are also small-world networks. This finding extends Gruenenfelder and Pisoni's result for English and sup-

[5]Following Gruenenfelder and Pisoni (2009), we estimate the ASPL of an Erdős-Rényi graph using the formula $ASPL_{ER} \approx \frac{ln(|V|)}{ln(\langle k \rangle)}$, where $\langle k \rangle = \frac{2|E|}{|V|}$ is the graph's average degree. The average clustering coefficient of an Erdős-Rényi graph is given by $C_{ER} = \frac{\langle k \rangle}{|V|}$ (Albert and Barabási, 2002).

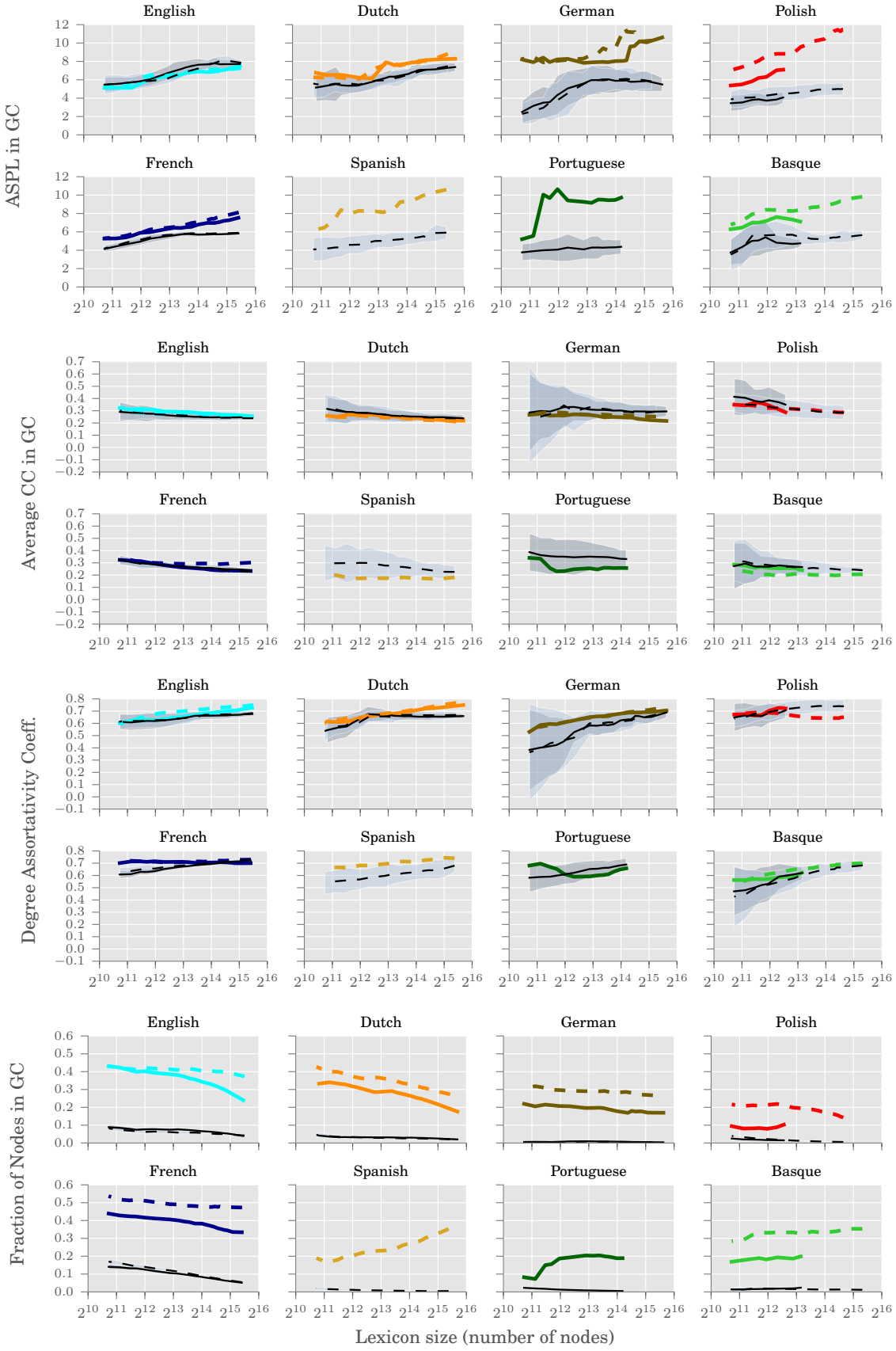

Figure 4: Thick, coloured lines show network statistics as a function of lexicon size in real lexicons. Thin, black lines show network statistics averaged across twenty random pseudolexicons, and the shaded regions indicate $\pm$ 2 standard deviations. Solid lines are for lemmas; dashed lines are for all wordforms.

ports their claim that the small-world properties arise naturally from the conventional definition of these networks, rather than suggesting a "deeper organization" as suggested by Arbesman et al.

Our results for degree assortativity also support Gruenenfelder and Pisoni's argument that substantial assortative mixing by degree can be expected in networks which are based on the overlap of parts randomly combined into wholes. The degree assortativity coefficients for all of the real lexicons lie between 0.5 and 0.8, which is higher than the values typically observed in social, biological, and technical networks. But again, the coefficients of the pseudolexicons are similar to those of the real lexicons, making these values less remarkable.

The final generalisation that Arbesman et al. made was that all the languages they examined had a smaller proportion of nodes in their giant components than the 80-90% that is typically observed in complex networks. We find the same, but as noted in Gruenenfelder and Pisoni's analysis of English, the fraction of nodes in the giant component of the real lexicons is actually much *greater* than in their matched pseudolexicons. This is not surprising, as the real lexicons have phonotactic constraints which would tend to make words more similar to one another than if phonemes were sampled at random. So again, the claim of a "deeper organisation" seems premature.

5.2.2 Cross-linguistic differences

The search for interesting universal properties is only half the motivation for making such cross-linguistic comparisons. Ideally, we also want to identify *differences* that might correlate with different behavioural patterns across languages. Two of our statistics don't reveal much on this point: average clustering coefficient (discussed above) and degree assortativity. There are some quantitative differences in degree assortativity between languages, but they seem mainly driven by differences in the word length distributions, since the differences across real lexicons pattern the same as the differences across pseudolexicons.

However, our results do reveal some more interesting cross-linguistic differences in the way the other two statistics vary across lexicon sizes and lexicon types (lemmas vs all wordforms).

Fraction of nodes in the giant component Some of the cross-linguistic differences in this statistic again seem driven by differences in word-length distribution, since there are large cross-linguistic differences for this statistic even in the pseudolexicons. For example, pseudolexicons matched to French or English tend to contain around 10% of the nodes, whereas pseudolexicons matched to German or Spanish are an order of magnitude smaller. However, word lengths cannot account for all of the differences in giant component size across languages, because the magnitude of the difference in values between the real lexicons and their corresponding pseudolexicons also varies across languages. For example, the giant component sizes of the pseudolexicons matched to Basque and Polish lemmas are reasonably similar, but the giant component sizes of the real Basque lemma lexicons are twice as large as those of the real Polish lemma lexicons.

In most of the languages the fraction of nodes in the giant component tends to decrease with increasing lexicon size (i.e. as more infrequent words are included in the lexicon), which suggests that less frequent words are phonotactically unusual. In contrast, in Spanish, Portuguese, and to a lesser extent Basque, the less frequent words are more likely to be a part of the giant component, suggesting that they are more similar to other words in the language. These different trends do not appear to be solely a consequence of differences in word-length distributions, since in the pseudolexicons matched to Spanish, Portuguese, and Basque, the fraction of nodes in the giant component does tend to decrease slightly with increasing lexicon size. This finding could be important for understanding cross-linguistic differences in language processing, since both word frequency and phonotactic probability are thought to affect both recognition and production of words (Luce and Pisoni, 1998; Vitevitch and Sommers, 2003).

These results also provide a real example of the behaviour hypothesized in Figure 2, underscoring the danger of using a single lexicon size to compare phonological network statistics across languages: if we compared lexicons containing around 10,000 words, we might conclude that Dutch wordforms were more densely connected than Spanish wordforms; whereas if we compared lexicons containing 30,000 words, their giant component sizes would support the opposite conclusion.

Average shortest path length Arbesman et al. noted that the ASPL for Mandarin was double that of Hawaiian, and raised the question of whether

this quantitative difference was significant. Our results suggest not: the sizes of the two lexicons they used (Hawaiian: 2578, Mandarin: 30,086) are similar to the smallest and largest sizes of the Polish and Spanish wordform lexicons used in our study (Polish: 1694 and 25,623, Spanish: 1694 and 42,461), and we see that for Polish and Spanish wordforms, as well as for Portuguese lemmas, the largest lexicon has almost twice the ASPL as the smallest one.

On the other hand, there do seem to be some meaningful differences in ASPL across languages. For the English lexicons, the ASPLs in the giant component are barely distinguishable from those of the corresponding random pseudolexicons. However, the values for Spanish and Polish lexicons are consistently higher than those of their respective pseudolexicons; while for German, Portuguese, and Basque, the differences between real and random lexicons are less stable across different lexicon sizes.

It should be noted that while the sizes of the pseudolexicons are matched to those of the real lexicons, the sizes of their giant components are not. Since the giant components of random pseudolexicons tend to be considerably smaller than those of real lexicons, it is unsurprising that their ASPLs tend also to be smaller. Nevertheless, our results show that the ASPL in the giant component of a phonological network is not a simple function of the giant component's size. Recall that the difference between the size of the giant component in the real lexicons and the size of the giant component in the corresponding random lexicons is smaller for Polish than for English or French. Hence, all else being equal, we would expect the difference in the ASPLs of real and random lexicons to be smaller for Polish too—but the ASPLs in the Polish giant components are actually larger, relative to the corresponding pseudolexicons, than those of English or French.

Polish also behaves differently from some of the other languages with respect to its morphology. In Polish, the magnitude of the difference in ASPL between real and random lexicons is greater when morphological variants are included than when the lexicons are restricted to lemmas, but this is not the case for English, Dutch, or French.

6 Conclusion

This paper has argued that, when making comparisons between phonological networks, researchers must consider that network statistics are affected by lexicon size, phoneme inventory size, the distribution of word lengths, and whether morphological variants are included or not. Since it is not possible to directly control for all of these in cross-linguistic comparisons, we have proposed that such comparisons need to be made indirectly, by looking at how each language's phonological network differs from a matched pseudolexicon across a range of lexicon sizes, and then comparing these differences across languages. While this approach doesn't permit simple comparisons of single numbers, nevertheless it can lead to insights regarding proposed universal properties as well as cross-linguistic differences.

In particular, our analysis of eight languages provides further support to Gruenenfelder and Pisoni's (2009) claim that the small-world and other properties discussed by Vitevitch (2008) and Arbesman et al. (2010) are a consequence of how phonological networks are defined, and do not necessarily reflect particular growth processes or cognitive pressures. At the same time, we did identify several differences in the behaviour of network statistics across different languages, which could provide an explanation for previously identified differences in language processing. We hope that our results will inspire further work to investigate these potential connections and to extend our analyses to additional languages.

7 Acknowledgements

This work was supported in part by a James S. McDonnell Foundation Scholar Award (#220020374) to Sharon Goldwater, and by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh.

References

Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47.

Samuel Arbesman, Steven H. Strogatz, and Michael S. Vitevitch. 2010. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03):679–685.

R Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. CELEX2, LDC96L14. *Web download, Linguistic Data Consortium, Philadelpha, PA*.

Maria Fernanda Bacelar do Nascimento. 2003. Um novo léxico de frequências do Português. *Revista Portuguesa de Filologia*, 25:341–358.

Matthew T Carlson, Max Bane, and Morgan Sonderegger. 2011. Global properties of the phonological networks in child and child-directed speech. In N Danis, K Mesh, and H Sung, editors, *Proceedings of the 35th Boston University Conference on Language Development*, volume 1, pages 97–109. Cascadilla Press Somerville, MA.

Matthew T Carlson, Morgan Sonderegger, and Max Bane. 2014. How children explore the phonological network in child-directed speech: A survival analysis of childrens first word productions. *Journal of Memory and Language*, 75:159–180.

Kit Ying Chan and Michael S Vitevitch. 2009. The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6):1934.

Susan Garrett, Tom Morton, and Cynthia McLemore. 1996. CALLHOME Spanish lexicon, LDC96L16. *Web download, Linguistic Data Consortium, Philadelpha, PA*.

Rutherford Goldstein and Michael S Vitevitch. 2014. The influence of clustering coefficient on word-learning: how groups of similar sounding words facilitate acquisition. *Frontiers in Psychology*, 5.

Inês Gomes and São Luís Castro. 2003. Porlex, a lexical database in European Portuguese. *Psychologica*, 32:91–108.

Thomas M Gruenenfelder and David B Pisoni. 2009. The lexical restructuring hypothesis and graph theoretic analyses of networks based on random lexicons. *Journal of Speech, Language, and Hearing Research*, 52(3):596–609.

Aric A Hagberg, Daniel A Schult, and Pieter J Swart. 2008. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA, August.

Trevor A Harley and Helen E Bown. 1998. What causes a tip-of-the-tongue state? evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, 89(1):151–174.

Jon Kleinberg. 2000. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 163–170. ACM.

Paul A Luce and David B Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1):1.

Paweł Mandera, Emmanuel Keuleers, Zofia Wodniecka, and Marc Brysbaert. 2014. Subtlex-pl: subtitle-based word frequency estimates for Polish. *Behavior Research Methods*, 47(2):471–483.

Boris New, Christophe Pallier, Ludovic Ferrand, and Rafael Matos. 2001. A lexical database for contemporary French on internet: Lexique. *Année Psychologique*, 101(3):447–462.

Mark EJ Newman. 2003. Mixing patterns in networks. *Physical Review E*, 67(2):026126.

Manuel Perea, Miriam Urkia, Colin J Davis, Ainhoa Agirre, Edurne Laseka, and Manuel Carreiras. 2006. E-hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, 38(4):610–615.

Rik Sarkar, Xianjin Zhu, and Jie Gao. 2013. Distributed and compact routing using spatial distributions in wireless sensor networks. *ACM Transactions on Sensor Networks (TOSN)*, 9(3):32.

Thomas Schank and Dorothea Wagner. 2004. *Approximating clustering-coefficient and transitivity*. Universität Karlsruhe, Fakultät für Informatik.

Tanja Schultz. 2002. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *INTERSPEECH*.

Massimo Stella and Markus Brede. 2015. Patterns in the English language: phonological networks, percolation and assembly models. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(5):P05006.

Michael S Vitevitch and Eva Rodríguez. 2005. Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders*, 3(1):64–73.

Michael S Vitevitch and Mitchell S Sommers. 2003. The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition*, 31(4):491–504.

Michael S Vitevitch and Melissa K Stamer. 2006. The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, 21(6):760–770.

Michael S Vitevitch. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(4):735.

Michael S Vitevitch. 2008. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, 51(2):408–422.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Johannes C Ziegler, Mathilde Muneaux, and Jonathan Grainger. 2003. Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, 48(4):779–793.